The Promise and Perils of Using Big Data in the Study of Corporate Networks: Problems, Diagnostics and Fixes

Eelke M. Heemskerk, Kevin Young, Frank W. Takes, Bruce Cronin, Javier Garcia-Bernardo, Vladimir Popov, W. Kindred Winecoff, Lasse Folke Henriksen and Audrey Laurin-Lamonthe

Revised version January 2017

Abstract

Network data on connections among corporate actors and entities – for instance through co-ownership ties or elite social networks – is increasingly available to researchers interested in probing many important questions related to the study of modern capitalism. We discuss the promise and perils of using Big Corporate Network Data (BCND) given the analytical challenges associated with the nature of the subject matter, variable data quality, and other problems associated with currently available data at this scale. We propose a standard process for how researchers can deal with BCND problems. While acknowledging that different research questions require different approaches to data quality, we offer a schematic platform that researchers can follow to make informed and intelligent decisions about BCND issues and address these issues through a specific work-flow procedure. Within each step in this procedure, we provide a set of best practices for how to identify, resolve, and minimize BCND problems that arise.

Keywords

Corporate Networks; Big Data; Network Data Quality; Diagnostics, Big Corporate Network Data.

1. *The Age of Big Corporate Network Data*

Corporations are tightly embedded in networks of power and control. Corporations share board members (creating interlocking directorates), they share owners, and they share holdings with one another. A sizeable literature has established that these networks facilitate the spread of corporate governance routines and practices from board to board through imitation and learning (among others Davis 1991; Haunschild 1993; Rao & Sivakumar 1999; Tuschke et al 2014). As a communication structure the network promotes the reproduction of existing beliefs and ideas, as well as the dissemination of new ones (Burris 2005, Mizruchi 1989; Carroll et al 2010). These networks have long formed distinct national business communities and have been part of the organization of national economies. Increasingly, however, these networks now transcend the national level and form a new complex global system of corporate ownership and control (Vitali et al 2011; Starrs 2013; Heemskerk & Takes 2016; Heemskerk et al 2016).

This fundamental reorganization of contemporary networks of corporate control has coincided with remarkable innovations in research practices. Over the last two decades the fields of computer science, physics, and complexity studies have become increasingly interested in complex network analysis, leading to a great number of breakthroughs in biology, sociology, finance and economics (Schweitzer et al, 2009; Borgatti et al, 2009; Barabasi & Albert, 1999; Battiston, 2016). At the same time new datasets are now available that allow us to start investigating standardized information on millions of firms and connections between them. Just a few years ago, scholars were manually collecting lists of 'Top 100' or 'Top 500' global firms through lists such as the Fortune 500 in order to evaluate the status of transnational elite ties (Carroll & Sapinsky 2002, 2010; Davis, Yoo & Baker 2003; Cronin 2012, Murray 2014). Studies of elite network community structure in particular regions, such as Europe, involved a few dozen (Van der Pijl, Holman & Raviv 2011), or a few hundred (Heemskerk, Daolio & Tomassini 2013; Carroll, Fennema & Heemskerk 2009) large firms. Today scholars have begun to scale their analysis to global

levels composing, for example, the largest 1 million firms in the world (Heemskerk & Takes 2016) or the .6 million most significant transnational corporations in a structure of global corporate control reduced from 30 million available firms (Vitali et al 2011). We call this scaling the emergence of Big Corporate Network Data (BCND).

This means that today we are able to combine advanced analytical and computational tools for analyzing big data on the one hand with theories on the architecture of the global economic order as a whole, on the other. Such studies are likely to proliferate in the years to come, raising new possibilities for research and new questions about the structure of contemporary capitalism (see Compston 2013). Complete, or quasi-complete, population studies are particularly promising for network analysts because datasets based on sampling limit the range of techniques and measures that one can soundly apply when conducting network analysis in particular (Marsden (1990); also see the debate on sampling issues in interlock studies by Carroll and Fennema (2004) and Kentor and Jang (2004; 2006)). More fundamentally, large-scale network data holds the promise to finally overcome the nagging boundary problem of network analysis. As Allen (1974: 396) stated in his pioneering work: "The most satisfactory sampling design for structural analysis is a saturation sample of the entire universe or population; however, this alternative is clearly not feasible for large social structures." Forty years later, we can confidently say that we have reached the phase where we can use big data to study the entire universe of interest.

Big Data, Big Problems?

While Big Data brings great promise, it can bring along Big Problems. Discussions associated with Big Data sometimes suggest that the sheer volume of data should reduce data quality worries (Mayer-Schönberger and Cukier 2013). Along this train of thought, missing observations and marginal inaccuracies are assumed to be washed away as error. While this is hardly correct for any kind of data since data is rarely missing completely at random, it is a particularly dangerous assumption to make in the context of network-relational data. That is because such missing data can significantly transform network topologies and thus observed network analysis results (Borgatti 2006, Mestres 2008). Some

network analytic measures and techniques are robust enough to reliably handle a few missing nodes or edges, but others, and often the more interesting ones, are highly fragile when faced with data incompleteness and sampling bias (Costenbader and Valente 2003).

At the same time there is a misunderstanding that the central challenge associated with Big Data, and potentially with Big Corporate Network Data, is only that of devising new computing architectures and algorithms (Jagadish 2015). It fuels the widespread perception that Big Data simply means scaling up of computational capacities and the development of new algorithms (see Agrawal et. al. 2014). We see the challenge of Big Corporate Network Data as presenting a set of *analytical* problems, and not simply technical ones. This is not to say that the volume does not change the researcher's relationship with the data. It does, and in significant ways. Utilizing Big Corporate Network Data sources from off-the-shelf information providers such as *Orbis*, *Boardex* or *Thompson One* essentially outsources the data collection. Whereas the manual hand-coding of the past was laborious, it provided the researcher with good grounding knowledge of the data. This intimate understanding of the data is now gone. This leads to a regular confrontation with BCND issues.

Our aim in this article is therefore not to present one specific technical fix, but rather to make a *meta-methodological intervention*. It represents the accretion of efforts from an international consortium of scholars from 12 different universities in 6 different countries. We came together after many bilateral conversations about how to address data quality in the context of the study of corporate elites. When searching for novel practice standards with our colleagues (for example, what to do with missing data in the context of corporate elite connections, or how to report entity resolution issues), we found we could not find any such standards of best practices. Based on our shared experience in dealing with BCND, we propose a standard process for what we consider to be the most appropriate way researchers should deal with BCND problems, acknowledging that different research questions require different approaches to data quality. Such standards are urgently needed so that scholars can more effectively measure what they seek to measure, so that they can compare alternative data sources, and ultimately so that scholars can better accumulate valuable knowledge about what corporate networks look like and how they may be changing. For these reasons it is

imperative to begin a conversation about research process standards *now* in order to advance the quality of the research community in the future.

In what follows below, we begin by sketching the problems that come with Big Corporate Network. We put forward a framework whereby we first separate the most fundamental issues with BCND in order to subsequently suggest a structured way to diagnose and fix these issues, using well-known characterizations. This takes the form of a schematic platform for making informed and intelligent decisions about BCND issues. These occur on multiple levels and involve different iterative steps, and thus we lay out a set of work-flow procedures that researchers can follow to address these issues, through a decision tree. Within each level of the decision tree, we provide a set of best practices for how to identify, resolve, or minimize BCND problems that arise. This means that while we do suggest methods to reduce uncertainty and noise from the data, our main goal is to be able to assess the extent to which data quality issues exist and what it means for the meaning that we derive from the analysis of concern. We introduce both new tools and techniques to diagnose the severity of BCND problems as well as specific techniques and fixes to deal with these problems.

This article is intended not only for researchers working on existing projects that confront BCND problems but also to encourage future scholars to engage with these data quality issues head on and through a systematic process rather than minimizing them. While the specifics of our recommendations can and will be adapted to different circumstances in future research, we also hope that reviewers of research use some of the insights we offer here to help improve the peer review process and in the interest of better science.

We do not take a position in the debate on the merits of a data driven vs. theory driven research, as we believe that the problems we discuss here are relevant for researchers in both domains. Our intervention is also not intended to be one specific to the study of corporate interlocks, although we do use it as an important running example.

We believe our suggestions extend wider than this kind of analysis, incorporating networks among corporations in general. While many existing studies have examined board interlocks among firms, recent analyses have extended to financial flows across firms (Battiston et al

2016; Squartini et al 2013), ties of ownership (Vitali, Glattfelder & Battiston 2011; Fichtner et al 2016; Haberly & Wojcik 2015), and other connections among elite interlocutors of firms that do not constitute board interlocks (Kim et al 2015). More generally we we acknowledge that the issues we encounter are paramount in other fields of inquiry related to network analysis as well. The suggested diagnostics and fixes may be applicable to these domains.

2. Big Corporate Network Data: Characteristics and Issues

The characteristics of Big Data are traditionally seen through the prism of 'three Vs': *Volume*, *Velocity* and *Variety* (Laney 2001). More recently additional V's have been suggested, including *Veracity* (Ward 2013), and *Variability* (Fan 2012). These V's provide us with a categorical context we can use to dissect the issues and problems we run into when working with BCND. In this section we therefore explore BCND through the lenses of these V's in order to determine the particular issues we need to address.

While *Volume* – indicating the sheer amount of data now available to researchers – is the most well-known characteristic of Big Data, we argue here that the volume *in itself* is not problematic in the case of BCND. A typical concern with the *Volume* of Big Data deals with the information processing challenges associated with data analytics (e.g. Fisher et. al. 2012). We do not focus on these technical issues because we see it as a misperception that the integration between Big Data and social science is about technical capacities. Certainly within the context of BCND the volume is larger than before, but manageable with current tools and techniques. However, the sheer *Volume* of the data does alter the relationship the researcher to the data, which in turn leads to a number of (analytical) issues that are related to the other V's.

First, BCND feature a *Variety* of information. Information is stored using different types of structured data and generally lacks universally employed unique identifiers. While the richness of these data is an asset, different data sources – or even the same data source at different points in time – may not use the same rules for collecting and coding data. One of

the key challenges confronting the study of large corporate networks is therefore *entity resolution* – the process of determining whether similarly named firms or similarly named individuals are the same or different actors. In addition, *Variety* means that data comparability and completeness may not be consistent across sets of data or different time points. Thus, it is increasingly important to know the mechanisms by which the data are collected, cleaned, and stored in addition to the data-generating process. Yet private information providers are not always keen in sharing this information. Another key challenge of BCND is therefore to assess the completeness of the data.

Second, BCND is characterized by *Velocity*. Traditionally, velocity refers to the fact that the flow of data is apart from massive, also continuous, constantly flowing in from different sources. BCND source databases are updated almost continuously, so the data is changing quickly as new information is added over time. This leads to new research opportunities, for instance utilizing longitudinal information. But it also means that some parts of the database may be updated while others are not. In the case of BCND we typically see that the more developed and the richer countries are, the better their corporate registries and hence the higher the *Velocity* of the data. This higher *Velocity* in some countries compared to others can lead to incorrect comparisons. In other words, the *Velocity* of BCND leads to the issue of *accuracy*.

Third, *Veracity* refers to the fact that the quality of data is often unclear. For instance, is the information on board composition correct and up-to-date? This is related to the issue of data provenance, which refers to the description of the origin, creation, and propagation process of data collecting (Glavic 2014) and the general logic of its extension and priorities. Data are collected through a variety of means and typically the precise collection protocols are not transparent and cannot be thoroughly audited. *Veracity* of BCND thus also leads to concerns about *accuracy* and *completeness*.

Finally, *Variability* refers to the fact that the way in which the user wants to interpret the data may change over time or according to research question. For example, in inter-firm networks we may sometimes be interested in studying firms with different corporate entities as one entity, whereas if we are primarily interested in the corporate structure we should

keep all the firm's legal entities distinct. *Variability* in the use of data requires us to understand how the data are constructed. But because of the *Variability* of BCND, it is crucial that the researcher is clear about its *unit of analysis*. What is it that you actually want to study? While this is obviously true for all studies, we argue that with big data in general and BCND in particular there is an increased risk of errors because data collection is not tailored to the research question. In practice, we often see that researchers devise research questions that try to utilize the full potential of new data sources. This is not problematic in itself, but it means that researchers may be tempted to use particular units or fields that are available in the data structure as objects as research. This can hold for both the nodes and the edges in the considered corporate network. It is therefore imperative to carefully consider if the BCND that is available does indeed correspond with the proper *unit of analysis*.

Some researchers consider Validity as yet another V of big data, referring to the question of whether the type of data that is considered, is suitable for measuring the considered phenomenon. For example, in the board interlock network, edges are often assumed to facilitate potential information exchange. Although we may be confident that the board interlock network correctly models the actual board composition, we do not necessarily know about the precise information exchange between the boards on a case-by-case basis. Also, different countries have different governance structures, rules and regulations. A non-executive director in China is not the same as a non-executive in the UK. A big data approach easily allows for study of, for instance, board interlocks across the globe, but decontextualizing boards and firms may lead to invalid conclusions. One way of seeing this is that validity refers to the veracity not of the data itself, but rather of the researcher's interpretation of the data (such as an edge) as a proxy measure for something else (information exchange). Therefore, it is essential that the researcher has a firm understanding of the theoretically informed unit of analysis. Given that potential problems, diagnostics and fixes for validity are similar to those of Veracity and Variability, we do not consider it separately in this article.

[INSERT TABLE 1 HERE]

Exploring the characteristics of Big Corporate Network Data brings us to four basic problems (see Table 1). These problems are not only relevant for Big Corporate Network Data. However, we argue that all studies that use BCND should carefully consider each of these questions: Are you clear about the appropriate unit of analysis? Is there entity ambiguity in your data? How complete are the data? How accurate are the data? These four questions may appear simplistic. However, reviewing the literature we find that typically studies do not report (sufficiently) on these issues. In part, this may be due to the above mentioned *idée fixe* that when one uses big data we need not worry much about data quality because the sheer volume of the data will counter the effect of missing or incorrect data values. And in part this lack of transparency on these basic questions may be related to the current deficiency of tools and techniques to assess the completeness and accuracy of the huge datasets we now use. To remedy this we propose a number of diagnostic routines and techniques for fixing data problems. These fixes are divided in two broad categories: *Semantic techniques* try to correct the diagnosed problems by using additional attribute information of the data, while *topological techniques* utilize network properties to assess and increase data quality.

Figure 1 provides a schematic overview of these four issues in the form of a decision tree. Proceeding through the decision tree, an honest answer can often be 'Not sure'. Therefore we also suggest a number of diagnostics to help the research community answer these questions. We hope that this decision tree and the suggested tools and techniques help researchers using corporate network analysis to more systematically answer important questions. Authors can increase transparency by providing an answer to these questions in their methods sections. The following section continues with a step-by-step discussion of each of these questions, diagnostics, and fixes as illustrated by the decision tree. These steps are sequential for a reason. The question about the unit of analysis determines what kind of data is going to be studied and selected from a source database, and represents an important conceptual step as one related to diagnosis of data quality. Entity ambiguity needs to be addressed before completeness, because incorrect entity resolution may lead to misleading statistics when completeness is assessed. Completeness should be addressed before accuracy, because based on the fact that certain segments of the data may be incomplete, we may wish to reduce the sample size to a complete segment or aspect of the data

[INSERT FIGURE 1 HERE]

3. Diagnostics and Fixes for Big Corporate Network Data

3.1 Step 1: Identifying Units of Analysis

3.1.1. Problems with Units of Analysis

When we pursue analysis of large-scale networks we can be tempted to simply consider the data within the dataset comprising the network of interest. But as students of corporate networks we must use a meaningful unit of analysis. This also means that we must have a clear definition of what constitutes a firm (node) in a given corporate network of interest and what constitutes an edge. With BCND this is not always a trivial task since corporations are composed of many interrelated legal entities. As Butts (2009: 416) remarked, "to represent an empirical phenomenon as a network is a theoretical act. It commits one to assumptions about what is interacting, the nature of that interaction, and the time scale on which that interaction takes place. Such assumptions are not 'free', and indeed they can be wrong. Whether studying protein interactions, sexual networks, or computer systems, the appropriate choice of representation is key to getting the correct result."

When approaching research questions related to corporate network data, one confronts a simple but important ontological question: what is a firm? While this question might be considered trivial for many kinds of analyses, for the study of corporate networks in particular it is a fundamental question about the definition of nodes and edges. Legal definitions matter because much data on corporate networks come from public registers. But as scholars we may not want to rely on lawyers' definitions of firms. Shell companies, for instance, disturb our common sense about what a firm is. Shell companies are legal entities without any underlying corporate activities and they are often set up to lower taxes (or, in more malign cases, to avoid corporate responsibility, liability, or to launder money). As such a board interlock between two shell companies is not theoretically equivalent to an interlock between firms engaged in actual corporate activities (see Heemskerk & Takes 2016). Furthermore, shell companies often have boards consisting mainly of lawyers and can have

formal board memberships in the hundreds or even thousands. These nodes fundamentally change the network topology in the corporate network of concern and leads to a careful reflection on whether we should consider shell companies as actors in our corporate network. This train of thought essentially feeds back to the initial basic question - what are the nodes and edges in our network and are they commensurable? - and is associated with the boundary specification problem in network research (Laumann et al. 1989; Carpenter et al. 2012).

Whenever we broaden our definition of edges or nodes, our network substantively changes its meaning and function. This is a central issue within network analysis (see Butts 2009). So even when the researcher has a clear understanding of what the nodes are, another boundary issue presents itself: what set of nodes and edges are part of the same network? This problem typically emerges when dealing with complete populations of firms in a given geographic context. Here it is advisable to question if any given population can meaningfully be thought of as *one network*. If we are interested in studying the Indian or the Dutch corporate network we sometimes want to qualify what this network consists of. Many studies for instance exclude wholly owned subsidiaries of foreign firms, such that IBM Netherlands is not considered part of the Dutch network (e.g. Stokman et al 1985). With small samples, the researcher can hand pick her sample. But this becomes a problem in large-scale databases if we can observe huge variations in the kinds of firms we have data about.

3.1.2 Unit of Analysis Diagnostics

There is no single diagnostic for examining if the network data well-represent the unit of analysis. We suggest an exploratory approach that takes into account several measures and reflects wisely on the research question of concern. The bottom line is to look for unexpected anomalies in the data. If we are interested in interpersonal networks based on affiliations, producing an appropriate plot of the distribution of affiliations among the population of individuals in the dataset is already likely to reveal anomalies in the data. Distributions of affiliations are highly likely to be long-tailed and any obvious spikes at the high end of the distribution could be an indication that an identifiable group of outliers is present in the data. Whether we then want to include this group of individuals or not is an

analytical question that should be clarified as we define (or re-define) our unit of analysis. In a similar vein we can look for deviances from structural characteristics in the data that a particular type of corporate network is known to generally display. If a core-periphery structure is usually found in a particular type of corporate network but is not so in a set of observed data, this could be caused by a systematic group of outliers that behave strangely (rather than that the actual network of interest does not have a core). If time-stamped data are available, it is possible to look for temporal anomalies. Using the raw data to plot how network-level measures of interest (e.g. centralization, cluster coefficients, core-ness etc.) vary over time can be useful here. If measures are volatile in ways that cannot be temporally explained (e.g. seasonality), we may want to check if alien groups enter the network of interest and disturb otherwise stable structural features.

3.1.3 Fixes for Unit of Analysis Problems

Two main ways of fixing the data problems raised above can be identified. A semantic approach is possible if we are able to locate a certain type of actor or edge as the root cause of our data anomaly – either from empirical knowledge about the network or from analysis of variance in node or edge attribute data. In that case we can make sense of our problem and make informed decisions about whether to exclude the source of the problem through targeted sampling. This approach is closer to what we might term ordinary data cleaning regardless if this work is aided by search and matching algorithms or done manually. A topological approach by contrast excludes certain nodes from the network of interest based on certain structural characteristics that such nodes display (such as degree) , and thus moves towards to more analytical-methodological end of the 'data cleaning and quality assurance'-spectrum. We illustrate how a combination of the two approaches can be useful in identifying, and dealing with, data anomalies.

Henriksen et al. (2016) study corporate networks of board members for the complete population of Danish firms 1990-2015. Their data set comprises 422,020 individuals, 208,417 boards and 1,677,688 board memberships with start and end dates of these memberships recorded. Building on Useem's (1984) work on corporate elites, the authors set out to apply dynamic K-core decomposition to understand the temporal evolution of the corporate 'inner circle' in Denmark. K-core decomposition works by recursively pruning

nodes with lower degrees and thus successively identifying subgraphs of increasing degree centrality (Batagelj and Zaversnik 2003). As the threshold for entering the successive subgraphs increases, the subgraph identified becomes ever more cohesive. Based on their detailed spell data they were able to create monthly time slices of the entire network and apply the same K-core decomposition procedure to all those time slices, in turn figuring out if the size and composition of the core was stable over time. Using this well-established method, it turned out that the composition of the core was highly unstable and its size varied tremendously.

What caused this instability was not however a fracturing of an 'inner circle' as found elsewhere (Chu and Davis 2016) but data anomalies such as those described above, where extreme degree values appear due to the presence of shell corporations. The method breaks down because shell companies form their own internal communities which are not densely connected with the true global center of the network. The degree of nodes within these communities is based on highly redundant ties within heavily overlapping boards. K-core decomposition is not well-suited to deal with such situations, and as noted above researchers are likely to come across entity ambiguity issues such as shell companies when they reach into Big Data territory through their investigations.

This situation can be corrected with the introduction of path-based centrality measures into the decomposition method. Introducing an additional threshold based on betweenness scores into the pruning process allows for such locally central K-cores to be ignored. Insofar as the interesting unit of analysis is a global core in a network, this method deals well with data quality issues such as the presence of shell corporations. Before introducing the betweenness decomposition method no stably convergent core could be identified, because coreness thresholds were overly affected by the highly central board members of shell corporations. After the introduction of betweenness into the pruning process a stable core emerges as can be seen in Figure 2.

Identifying the problem, why it was a problem and how to fix it required an exploratory use of both the semantic and the topological approach, where defining the unit of analysis and the population of a network is part of the process of analytical discovery, relying in part on

familiarity with network analytic tools to understand topological characteristics and in part on more simple methods of finding data anomalies such as sampling and checking semantics.

[INSERT FIGURE 2 HERE]

3.2 Examining Entity Ambiguity

3.2.1 What is the Problem of Entity Ambiguity?

Low quality and integrity of corporate network data poses a fundamental threat to the validity of inferences drawn with Big Data. A simple example illustrates this point. Consider the following example, which utilizes data from *Boardex*. To investigate connections between public authorities and large global firms, researchers took the first-and-second degree connections from just three significant financial regulatory authorities in the North Atlantic: the Bank of England, the US Federal Reserve Board, and the US Securities and Exchange Commission. These public entities are highlighted in green within the network in Figure 3a. Also highlighted, however, is 'Goldman Sachs'. Yet as one can see, there are actually not one Goldman Sachs but rather 5. The centrality of Goldman Sachs in this network is unknown; if one wanted to know the connections between Goldman Sachs and these selected public entities not only would there be clear biases in the data but there would be 5 different measures generated for each. This kind of problem with entity resolution will bias measures of network structure, and the problem will only grow more severe with the expansion in size of the network. In the context of traditional datasets of a few dozen or hundred firms in a network this may not be a significant problem, as duplicate entities can be resolved efficiently and comprehensively through manual checking or sorting. In a Big Data context it is not feasible to do this comprehensively. Figure 3b shows an example of a 'resolved' network (see Marple et al 2017), in which not only Goldman Sachs but many other entities in the network have been resolved, generating significant changes in network structure and revealing the more genuine location of Goldman Sachs within the network.

[INSERT FIGURE 3 HERE]

One keystone form of entity resolution in a Big Data context is to use string matching algorithms. String matching algorithms: essentially a procedure in which one compares all entries in the data with all other entries in the data, and computes a similarity score for each pair of data entries. Then, the most similar pairs (i.e., with the highest score) can be deemed identical and subject to replacement. In the above example the two Goldman Sachs' can be reduced to one, and the subsequently network rewired to ensure greater accuracy. Scholars have a variety of string matching algorithms at their disposal (discussed below), which often entail measures of similarity across entity names. Yet even string matching only works with a degree of confidence; given the absence of manual checking the confidence intervals being relied on need to be transparent (see Takes and Garcia-Bernardo 2016).

Yet not all entity resolution issues are this simple. Many firms are part of highly complex corporate ownership structures which compound entity resolution challenges. According to the *LexisNexis Corporate Affiliations* database, for example, Bank of America is composed of 229 different legal entities, including subsidiaries and shell companies. The nature of such corporate hierarchies is likely to generate biases in network structure if left uncorrected. To illustrate the complexity of corporate forms Figure 4 provides a network representation of two large global corporations, Citigroup and Exxon Mobil. With the global parent in the center one can represent each successive level of the firm, from subsidiaries, separate holding companies and shell companies that run through the corporate hierarchy of the parent. The blue dots in the network represent legal entities that have the name stem of the global parent in their name (e.g. 'Citi' or 'Exxon') and which could be potentially resolved through string matching. The red dots however represent legal entities that do not have the name stems of the parent in their name. Exxon Mobil example, contains entities in its corporate hierarchy such as 'Houghton Realty Trust', which branches out from 'XTO Energy', which is a subsidiary of Exxon Mobil.

[INSERT FIGURE 4 HERE]

3.2.2 Diagnosing Entity Ambiguity

We recommend running simple diagnostics in the event that a researcher is unsure of the need to resolve entities. These can come in the form of simple string matching algorithms

that, for example, search for all variations of the name 'Met Life' and replace accordingly. Note that this will only identify certain kinds of entity ambiguity issues. As such it may make sense to gather auxiliary information on the structure of a given corporate hierarchy (such as ownership data) for a firm that is prominent in the data, and then run automated string search algorithms for all firms in a complex corporate hierarchy structure.

Corporate hierarchies not only include subsidiaries of a holding or parent firm, but also specialized holding companies and non-operating entities (otherwise known as shell companies). However, it is possible to diagnose the best way to approach resolving the data given the constraints a researcher faces and given the research questions that they are pursuing. Figure 5a below shows a scatterplot of the (ln) number of recorded subsidiaries of the largest half million firms in the world, on a global ultimate owner (or GUO) basis, using total assets as the indicator of firm size and using data from *Orbis*. The slope of this relationship reveals that the larger the firm, the more subsidiaries the firm is likely to have. Figure 5b shows a more select example with less off-the-shelf data. Because subsidiaries are only one form of entity, we counted the length of corporate hierarchies across the universe of corporations by randomly sampling 100 entities from 4 different strata of firm size (from ranks 1 to 500, 500-1000, 1000-10000 and 10000-25000) within the largest 25,000 firms across 59 countries from *Orbis*. For each of the 100 firms sampled we looked up their detailed corporate hierarchy information including subsidiaries, branches, units, holding companies, and non-operating entities related to the global parent. For each distinct entity we counted an additional unit of length in the global parent's corporate hierarchy (see Marple et al 2017). The regularity found within the data, illustrated in Figure 5b, is largely the same as for subsidiary data described above. The larger the firm (ranked in terms of assets), the more entities there were within a firm's corporate hierarchy. In the global distribution of firms, it is the 'top end' of firms that have the longest hierarchies. Such empirical regularities tell us which kinds of firms are likely to have 'longer' corporate hierarchies than others. Thus it helps to narrow the focus of which kinds of firms to focus on in terms of entity ambiguity fixes when working with BCND research questions.

[INSERT FIGURE 5 HERE]

Entity ambiguity can also be diagnosed through measures of network structure itself, i.e., the topological approach. For example, in an unresolved network of director interlocks among financial firms in the US, a researcher will quickly find numerous director ties between for instance 'Bank of America NA' and 'Bank of America Securities'. Thus in some instances entity ambiguity may be identified through abnormalities in tie structure. One way to identify such abnormalities may be through plotting a diagnostic of edge width across a modeled network. For example, plotting the network using strong attraction between connected nodes and strong repulsion among unconnected nodes can show clusters of highly connected nodes (See Figure 6). These nodes are sometimes local branches of small companies, which can be joined together or excluded from the sample. This method is explained in detail by Garcia-Bernardo and Takes (2016), and apart from visual inspection also proposes a number of topological network metrics that can be used to characterize such dense clusters.

[Insert Figure 6 about here]

3.2.3 How to Deal with Entity Ambiguity

In order to address entity ambiguity problems, the unit of analysis may require merging together related nodes. For example, we may be interested in the relationships between corporations, and thus would like to merge together all firms involved in the corporate structure. The best approach here is a topological approach, in which we use information about ownership to merge related companies. However, we do not always have ownership information. In those cases, we can merge companies that cluster tightly together (topological approach, see Figure 6 above), or by using firm names or other firm attributes (semantic approach). The entity ambiguity problem still persists when similar company names correspond to different companies (e.g. 'ASN' and 'ABN'), and when different names are part of the same corporate structure (eg. 'Zao Master D' and 'Beta Properties INC' are part of 'METLIFE INC'). It is of course possible to utilize this kind of information when it is available. Marple et al (2017) and Young et al (2017) utilize large lists of corporate family structures, including branches, subsidiaries and shell companies, among the largest 500 corporations in the world as the basis to batch-replace existing names in the network.

Utilizing string matching in all these processes is quite crucial, whereby two nodes are merged if their name is similar, thus re-wiring the network. Similarity can be measured in terms of the effort that it takes to convert one string into the other by modifying individual characters (edit-based measures), in terms of the number of words or n-grams that are shared between the strings (token-based measures), or a combination of both (hybrid measures) (Cohen et al, 2003; Bilenko et al, 2003). Since different variants of a company name usually differ in the ending (e.g. 'Bank of China limited' and 'Bank of China LTD'), algorithms that give lower weights to the end of strings are preferred. This is the case in the Jaro-Winkler (edit distance) algorithm (Winkler, 1990), and the term frequency–inverse document frequency (TF-IDF) using n-grams (token distance) algorithm (Salton et al 1975). Metadata information can provide useful supplements for entity matching. If two companies share the address and have similar names, it is likely that it is the same company. Moreover, several similarity values can be combined using machine learning algorithms such as neural networks (Cohen et al, 2003; Bilenko et al, 2003).

Given size and computational capacities for string matching in extremely large adjacency matrices, string matching can be performed on weighty edges above a given reported (high-in-distribution) threshold, where false ties are likely to exist (for example MetLife Inc has 200 connections to MetLife, etc) because of the nature of large corporate groups. Another way to exploit network structure as part of an entity ambiguity fix is to utilize community clustering. Marple et. al (2017) use community detection algorithms to separate node names into clusters, which are then sub-processed using string matching methods within each cluster to ensure greater accuracy in name replacement.

Because entity ambiguity problems can be highly complex, multiple methods might be used but in each case researchers should report entity ambiguity statistics – such as the frequency of name replacements, and if possible the precision-recall estimates associated with some form of 'ground-truthed' subset of the data. Such a subset can be a sample but can be a useful metric for how well name suggested name replacements perform. If the sample is large enough and the entity ambiguity problem significant enough (or *unknown* enough), the performance of each entity ambiguity fix can be reported through forms of measurement developed within the computer science community that measure precision and recall

performance of a given method. Precision, or positive predictive value, is the faction of retrieved instances that are relevant, while recall, otherwise known as 'sensitivity', is the fraction of relevant instances that are retrieved. Synthetic scores exist of precision and recall performance that researchers using BCND data can utilize, such as F1 scores and measuring the area under the ROC curve, which can help a researcher decide which entity ambiguity fixes are generating the best performance. An ROC (or Receiving Operator Characteristic) curve is a statistical metric used to visualize the performance of any classifier with two possible outcomes. It represents a plotting of the true positive rate against the false positive rate to understand the tradeoffs between sensitivity and specificity of a given classifier. An F1 score is another way to assess the performance of a classifier, as it measures the combined performance of precision (the fraction of retrieved instances that are relevant) and recall (the fraction of relevant instances that are retrieved). For an F1 score to be high, both precision and recall should be high. These should be used in light of what earlier diagnostics suggested, and the potential sensitivity of the network-relational measures that are ultimately being pursued. Treating entity ambiguity seriously and systematically not only facilitates the goal of analytic transparency and hence reproducibility but also will provide other researchers with information about error rates and the severity of entity ambiguity issues for specific research problems with large off-the-shelf Big Data datasets.

3.3 Completeness

3.3.1 Problems of Data Completeness

Data incompleteness can affect the trustworthiness of the inferences that made from statistical models of relationships in the data (Rubin 1976). Missing data in a network context is generally more problematic than it is in non-network contexts because of interdependencies in network data (Borgatti et al 2006; Kossinets 2006). Missing information in networks can have a multiplicative effect: each missing link directly affects two nodes and indirectly affects potentially many others. Completeness affects centrality scores, community detection algorithms, and comparisons between networks (Žnidaršič et al 2012). All of these metrics are critical for descriptive or inferential analyses of networks.

There are three main types of data missingness in networks: omission of actors and/or affiliations that actually exist in the network due to boundary specification, non-responsiveness to surveys used to construct the network data or an inability to construct a full network from observational data, and censoring according to a node-level characteristic such as size or prominence (Kossinets 2006). There are several reasons why some data may be missing, several of which are most relevant for our purposes. First, information providers usually collect complete data for large companies and at-least partial data for medium-sized companies, but may not report even the existence of many smaller-sized companies. For example, the database *Orbis* has complete information about Greek companies with greater than 250 employees but only contains around 2% of the companies with fewer than 10 employees (see Figure 7). While it is true that small companies are often less significant in corporate networks than large companies, ignoring the missing data could still significantly alter our results. In addition to differences in filling requirements, data completeness is higher in developed economies than poor economies and tax havens. Second, some data, particularly those describing interdependencies such as ownership relationships or corporate interlocks, may be collected egocentrically, and parts of the network will appear less connected than others as a consequence of the sampling procedure.

[INSERT FIGURE 7 HERE]

3.3.2 Completeness Diagnostics

Missing information can refer to the nodes themselves (e.g. missing companies or people), to the edges (e.g. missing director positions or ownership relations) or to metadata (e.g. financial information). Moreover, the data can be missing completely at random (MCAR), or missing with a probability that depends on an observed variable (missing at random, or MAR) or an unobserved variable (missing not at random, or MNAR). Missing metadata is usually correlated with another observed variable – for instance it is more likely that we lack data on firm assets if we are also missing data about the revenue of that firm. Because the variables are related, missing metadata can be imputed as long as some of the metadata is observed. Deleting cases under MAR is not recommended since it produces large biases (Rubin 1987).

Missing nodes or edges are commonly correlated with an unobserved variable, and are thus harder to study because imputation is not likely to reduce bias or inefficiency. Missing nodes can alter the results significantly. For instance, if we would want to analyze the characteristics of the agriculture sector using data from Orbis, we would find out that the average Mexican company is larger than the average US company. But this is due only to the better recording of small companies in the US. Because our results can be erroneous if there are missing nodes, it is paramount to assess the completeness of the data. Completeness diagnosis consists of the comparison of the data (or a subset of the data) to an external database that is known to be complete. Importantly, this step will often require aggregating the data by sector, country, or type of company. Figure 7 shows the completeness of the Greek data from the Orbis database by comparing it to Eurostat data. This step provides a first assessment of the type of data that is missing – small companies. If the first step reveals that we have missing data, a finer characterization of the missing data is needed. In this second step we look for the pattern of the missing data. For example the distribution of the majority of firm economic indicators, such as operating revenue, assets, or number of employees follow lognormal distributions. By comparing our database to external databases we can characterize the distribution of the missing data.

Although missing edges do not affect network measures as strongly as missing nodes, they can still affect the analysis (Kossinets, 2006). Diagnostics also require comparing our data with a complete database. Because complete databases of edges are not readily available, we usually rely on manual checks of a sample of the data. These manual checks in our experience usually show that while big companies have complete information about directors and ownership relationships, small companies lack such information, and thus have more missing edges.

3.3.3 Addressing Data Completeness

Once we have a clear understanding of the type of missingness in the corporate data we can deal with the problem in two ways: either by restricting our analysis to a part of the network with good quality data, hoping that the part missing does not bias inferences taken from the part we observe; or we can seek to improve the quality of the data to mitigate the effects of missingness. If we choose the latter there are two basic ways to operate under conditions of

incomplete data: an approach based on leveraging the assumptions regarding sampling procedures, and an approach based on imputation of estimated data in place of the missing data. If we choose to impute data the process should be done transparently and, if possible, repeatedly.

The simplest missing data to correct are metadata (e.g. the attributes of a firm or individual). Unless the amount of missingness is large these can be imputed using normal statistical procedures in the tradition of Little and Rubin (1987), including modern implementations such as multiple imputation and hot-decking (Cranmer and Gill 2013, Blackwell et al 2015a, Blackwell et al 2015b). These provably reduce biases and improve efficiency when data are missing at random.

Non-metadata missingness can exist at the node-level or tie-level. These are more difficult to impute than metadata, but some reasonable strategies have been developed. Information missing randomly at the tie-level (e.g. through representative sampling) is the more straightforward of the two, and can be imputed by inference using the latent space positions of nodes to replicate missing edges (Ward et al 2003). This approach might be useful under conditions of representative sampling or egocentric analysis. Huisman (2014) notes, however, that "simple" single-imputation methods, i.e. *ad hoc* methods that do not involve multiple imputation, are frequently biased. Until very recently this left few options for scholars working with missing network data other than deletion.

The most flexible type of imputation strategy for missingness, either at the node-level or edge-level, involves estimation of missing values using the likelihood-based exponential random graph model (ERGM) family (Handcock and Gile 2010), with extensions for Bayesian "data augmentation" (Koskinen et al 2010, 2013). In these models, expected values are imputed for missing data during the estimation of the ERGM parameters, and because ERGMs are estimated via simulation (usually employing standard Markov chain Monte Carlo methods) the missing data are imputed many times. While still quite new and rare these ERGM-based procedures have been successfully implemented on real-world network data suffering from quite complicated patterns of missingness (Wang et al 2016), and have been mathematically extended to temporal ERGMs (Ouzienko and Obradovic 2014). More

validation is needed to fully understand the properties and reliability of these model-based methods, but they possess considerable promise for scholars analyzing network data containing missing information.

What is known for certain is that data missingness presents serious problems in a network context even if the data are missing at random. Scholars should make every effort to correct bias that could emerge from missingness or, at minimum, to understand it's likely effects.

3.4 Accuracy

3.4.1 Problems of Data Accuracy

In general, accuracy refers to whether a measurement of data conforms to the real world. Specifically, we are concerned with whether the data consistently and correctly matches our conceptual understanding of corporate networks. Because corporate data is typically gathered in a rather indirect route via annual reports, business organizations, and (intermediary) information providers, and is furthermore changing over time (respectively the veracity and velocity aspects of big data), accuracy can be a significant issue. It affects whether the existence of a node or a link is correct and current. Accuracy is furthermore related to the question of whether the observed data accurately represents the type of network structure we are interested in studying.

An example illustrates how inaccurate corporate data can lead to an incorrect corporate network. We create a board interlock network from data on firms and directors in Panama, modeling social ties between firms that share board members. The data provider (Bureau van Dijk's *Orbis* database) aggregates data from different information providers in different countries, and is thus dependent on these data providers for the quality of its delivered content. In *Orbis* there are a total of 841,487 active firms for Panama. This large number suggests that a significant proportion of the firms in Panama have been collected. With the purpose of studying the Panama board interlock network, we selected for each of the 628,289 firms that actually had information on its board composition, the senior directors of people listed as currently holding a position at a particular firm. This yielded 3,172,041

unique director positions. In total, there were 1,207,541 unique directors. Given the significant number of board interlocks and a large number of directors, we may feel that we are on the right track to extracting a sensible network based on interlocking directorates. When we further inspect this data, for example by looking at the average board size -- which is 3,172,041 / 628,289 = 5.05 -- the data still seems accurate. Even the distribution of the number of directors per board seems sound, as shown in Figure 8, with an average around 5 and a few far less frequent larger boards (note the logarithmic vertical axis). When we examine the average number of positions held by a director -- which is 3,172,041 / 1,207,541 = 2.62 positions -- there is still no reason for alarm. In fact, it suggests an exciting number of interlocks.

[INSERT FIGURE 8 HERE]

However, when we look at the distribution of the number of positions held by a director in Figure 9, we see something alarming: in Panama there are directors with extremely large numbers of positions, led by one director in our data holding 16,744 positions at different firms. The names of these directors in the tail of the distribution are not common names that are wrongly matched by name matching or entity resolution software, but the majority appear to be actual unique names of directors with an enormous number of positions.

It is clear that including these directors is not beneficial to studying the board interlock network, for at least two reasons. First, from a theoretical (semantic) point of view it is unlikely that this person is actually facilitating a number of interlocks with clear causes and consequences for Panama's corporate structure. From a network-topological point of view, if we study the firm-by-firm network of Panama, then this director alone would create (16,744 * 16,743) / 2 = 140,172,396 interlocks as part of a fully connected clique of firms. Clearly, this is beyond the scale of any meaningful corporate network, especially given the fact that there are almost 100 directors with such a large number of positions.

Second, it should be noted here that the projection from the two-mode board-director network to the one-mode board interlock network is responsible for the quadratic increase in the number of interlocks, a general problem that arises when projecting two-mode

networks with skewed degree distributions (see Neal, 2014). Yet most likely, the example above is a by-product of Panama's now well-publicized status as a tax haven and host to large numbers of shell companies established for the purpose of evasion. Indeed, when we compare the distribution of the number of positions per director with other countries, such as Finland, Denmark, the United States, Netherlands and Sweden, as shown in Figure 9, the long tail in the Panama distribution can be compared with other states.

[INSERT FIGURE 9 HERE]

3.4.2 Accuracy Diagnostics

One obvious way of detecting the issues described above is by doing an analysis of the distribution of the data points. In particular, looking beyond mere totals and averages, searching for outliers in terms of frequency and value (thus, looking at the vertical and horizontal extremes and the outliers of this distribution) may allow one to catch errors or problematic idiosyncrasies in the data rather easily. More specifically, if we were generating the global board interlock network, we would generate these distributions for each of the natural groupings of firms into countries, and see if the distribution for each country makes sense. Ultimately, without manually inspecting the millions of director positions, we are able to automatically obtain insights in outliers. Of course then, one can manually look these outliers up in in the data to determine the actual reason.

3.4.3 Addressing Data Accuracy

One way of solving the issues pointed out in the example above is by setting sensible filtering thresholds based on what is known about the data. For example, suppose a researcher finds a given corporate board member with many more ties to other firms than average. In such instances the decisions taken to exclude or filter data should be transparently reported. Some authors on the current paper have for example encountered corporate network data in which there are extremely dense ties between two organizations, which were discovered to be the same organization, named differently, and thus a kind of entity ambiguity problem (see Section 3.2). Others have encountered instances whereby some individuals or firms possess exceptionally more ties to others than the rest of the

distribution. In these latter instances, these exceptional linkers/firms have been investigated and found to be particularly elaborate shell company structures (Henriksen et al. 2016). Indeed, the presence of such administrative ties (rather than social ties) between firms is often encountered (Heemskerk and Takes 2016; Takes and Heemskerk 2016). One solution is to take the presence of such administrative ties into account when interpreting the results. A second solution is filtering data to exclude such formations. Regardless, in every case the precise extent and steps should be reported. This is especially important given that normal possibilities of replication for BCND data are often not available given the proprietary nature of the data *and* the fact that BCND data purveyors are often updating their bespoke datasets in real time.

More generally, when data is so large that assessing its quality cannot be done by hand, more reliance on quantitative inspection is needed – examination of totals, averages, and distributions – rather than qualitative inspection. These can be compared to extant studies of a similar regulatory context and can be compared over time to check for the presence of major discrepancies in data accuracy arising from data volatility. Quantitative inspection also helps to segregate the data and compare the different quantities across different segregations of the data to detect outliers. For example, by segregating the data illustrated in Fiture 9 above by country we find significant outliers in Panama, which we know from the so-called 'Panama Papers' and other sources to be home to a great number of shell companies and tax-evasion entities. It is worth asking whether the prevalence of shell companies is accurately capturing the concept of interest, which may be corporations that are active in some type of production. If that is the case then scholars may need to prune the data according to some characteristic -- an above-zero number of employees, for instance, or some output-based characteristic -- so that the prevalence of shell companies does not distort the inferences that can be made from these data.

There is no panacea to these issues, no perfect statistical 'check' for data accuracy. Correcting data that inaccurately captures the corporate structures of interest requires wisdom and patience, and should therefore be done transparently. Data inaccuracy may also be related to some other issue with BCND – in particular data completeness and entity

resolution – so a holistic strategy to ensure the data is in proper condition for the analysis to be performed is desirable whenever feasible.

4. Discussion: Toward Common Goods

In this paper we have sought to add productive fuel to the conversation over data quality when utilizing what we have called Big Corporate Network Data (BCND) problems. Even prior to the advent of data on this scale there has been a rich discussion regarding how to best study elite networks (e.g. see Carroll and Fennema 2004; Kentor and Jang 2004, 2006). These and other longstanding issues within this specialized literature (among them what a given edge-relationship actually 'does' – see Cronin 2011; Mizruchi 1996) do not go away. They simply get compounded and added to a litany of other research challenges.

We have advanced a framework to help guide not only individual researchers but future discussions among the research community when it comes to data quality and the means of addressing them. Researchers should identify whether the data matches the unit of analysis, they should address entity ambiguity, data completeness, and data accuracy – and they should report on these steps, and their preferred diagnostics and fixes. We introduced both new tools and techniques to diagnose the severity of BCND problems as well as specific techniques and 'fixes to deal with these problems. Specifically within each level of Figure 1, we provided a set of best practices that we are aware for how to identify, resolve, or minimize BCND problems that are known to arise.

Our contention is that the research community would greatly benefit from walking through the flowchart proposed in Figure 1, or something close to it, and then transparently reporting about each step is a good recommendation for forthcoming research about corporate networks. Each step of the decision tree is supported by a variety of diagnostic tools for the unit of analysis, entity ambiguity, data completeness and data accuracy, as summarized in Table 2. As discussed in section 3, there are many existing diagnostic tools that can be deployed to improve the rigor of BCND; extensive and transparent use of these is recommended. Yet, while these provide a standard set of metrics that allow interpretation of the validity of analysis, the context of the metrics remains ambiguous as little is yet known

about the typical distributions of these metrics in standard corporate settings. This opens a research agenda for further development of these diagnostic tools.

[INSERT TABLE 2 HERE]

This list of suggested diagnostics and fixes will no doubt change and improve over time. While our intervention is geared towards the community of scholars working on corporate networks, we acknowledge that the issues we encounter are paramount in other fields of inquiry related to network analysis as well. The suggested diagnostics and fixes may be applicable to these domains.

To be clear, we are not claiming that work that does not give an indisputable answer to the proposed set of questions should remain unpublished. However, similar to caution exercised when drawing conclusions from correlations with large variance, we call upon researchers to exhibit significant awareness when interpreting corporate network analysis results when this data is subject to issues around data completeness and accuracy.

Assessment of the validity of BCND analyses in the future would be greatly enhanced with better knowledge of standard distributions of nodes, edges, financial variables and typical network structures in various corporate settings. These vary by sector, company type and country because of the differing competitive and regulatory imperatives in these settings. Because of the great impact of entity ambiguity on the shape of corporate networks, and consequent validity of any analysis, there is also a particular need for greater standardization of entity disambiguation methods. There will be great benefits from the development of standardized corporate entity matching algorithms with explicit confidence intervals. Further progress in the development of databases with unique identifiers and positions in corporate hierarchy will also aid this process.

Ultimately, the goal of a Big Data approach is to extract Value. For us, this translates to knowledge. A key question remains whether or not a Big Data approaches to questions related to corporate networks provides additional insight compared to studying small data. Indeed, even if one effectively avoids all the potential problems with BCND discussed here,

lacking a compelling justification for undertaking the analysis is still problematic. The mere fact that the availability of big data means a particular analysis *can* be conducted is in itself not a sufficient justification that it *should* be conducted. Ultimately, the proof of the pudding is in the eating, and with this intervention we hope to contribute to a vivid, candid and critical academic debate on the merits and pitfalls of using big corporate network data. In our view, early studies utilizing big corporate network data already led to revealing insights, for instance on the high level of concentration of global corporate control (Vitali et al 2011); on the hitherto disregarded multilevel nature of board interlock networks (Heemskerk et al 2016), and on the unprecedented shareholder power position of the Big Three passive investors in global equity markets (Fichtner et al. 2016). The promise of Big Corporate Network data however goes well beyond these arguably rather descriptive contributions. Crucial next steps include understanding the driving forces behind network dynamics by utilizing advanced modeling frameworks for big data, and ultimately pinpointing the economic, political and societal consequences of the newly uncovered patterns. These contributions cannot be made systematically without first addressing key challenges associated with BCND problems.

References

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., and Widom, J. (2014) 'Challenges and Opportunities with Big Data, A Community White Paper,' *Mimeo*, available at: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf.

Barabási A. and Albert R. (1999) 'Emergence of scaling in random networks,' *Science*, 286(5439), 509-12.

Batagelj, V. and M. Zaversnik (2003) 'An O(m) algorithm for cores decomposition of networks,' *arXiv*, preprint cs/0310049.

Battiston, S., Farmer, J. D., Flache, A., Garlaschelli, D., Haldane, A. G., Heesterbeek, H., ... & Scheffer, M. (2016). Complexity theory and financial regulation. Science, 351(6275), 818-819.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., Fienberg, S. (2003) 'Adaptive name matching in information integration,' *Ieee Intelligent Systems*, 18 (5), 16-23, doi: 10.1109/MIS.2003.1234765.

Blackwell, M., J. Honaker and G. King. (2015a) 'A Unified Approach to Measurement Error and Missing Data: Overview and Applications,' *Sociological Methods and Research*. Forthcoming.

Blackwell, M., J. Honaker and G. King. (2015b) 'A Unified Approach to Measurement Error and Missing Data: Details and Extensions,' *Sociological Methods and Research* Forthcoming.

Borgatti, S.P., Mehra, A., Brass, D., Labianca, G. (2009) 'Network Analysis in the Social Sciences,' *Science*, 323 (5916), 892 – 895.

Borgatti, S.P., Carley, K.., and Krackhardt, D. (2006) 'On the Robustness of Centrality Measures under Conditions of Imperfect Data,' *Social Networks*, 28 (2), 124–136.

Burris, V. (2005) 'Interlocking Directorates and Political Cohesion among Corporate Elites,' *American Journal of Sociology* 111(1), 249-283.

Butts, C. 2009. "Revisiting the Foundations of Network Analysis", *Science* 325(5939): 414-416.

Carpenter, M.A., Mingxiang, Li, and H. Jiang (2012) 'Social network research in organizational contexts a systematic review of methodological issues and choices,' Journal of *Management*, 38(4), 1328-1361.

Carroll, W.K. and M. Fennema (2002) 'Is there a transnational business community?' *International Sociology*, 17(3), 393-419.

Carroll, W.K. and M. Fennema (2004) 'Problems in the study of the transnational business Community,' *International Sociology*, 19(3), 369-378.

Carroll, W.K. and J.P. Sapinski (2010) 'The global corporate elite and the transnational Policy Planning network, 1996-2006 A structural analysis,' *International Sociology*, 25(4), 501-538.

Carroll, W.K., Fennema, M., and E.M. Heemskerk (2010) 'Constituting Corporate Europe: A Study of Elite Social Organization,' Antipode 42(4), 811-843.

Chu, Johan and Davis, Gerald. 2016. "Who Killed the Inner Circle? The Decline of the American Corporate Interlock Network", *American Journal of Sociology* 122(3): 714-754.

Cohen, W. W., Ravikumar, Pradeep D., Fienberg, Stephen E. (2003) 'A comparison of string distance metrics for name-matching tasks,' *Proceedings of IJCAI-03 Workshop on Information Integration*, 73--78, August.

Compston, H. (2013) 'The network of global corporate control: Implications for public policy,' *Business and Politics*, 15(3), 357-379.

Costenbader, E. and T.W. Valente (2003) 'The stability of centrality measures when networks are sampled,' *Social Networks*, 25(4), 283-307.

Cranmer, S.J. and Gill, J.M.. (2013) 'We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data,' *British Journal of Political Science* 43:2 (425-449).

Cronin, B. (2011) 'Networks of corporate power revisited,' *Procedia – Social and Behavioral Sciences* 10(5), 43-51.

Cronin, B. (2012) 'Transnational and national structuring of the British corporate elite,' in G. Murray and J. Scott (eds.) *Financial Elites and Transnational Business : Who Rules the World?* Edward Elgar Publishing Limited: Cheltenham, 177-92.

Davis, G. F. (1991) 'Agents without Principles? The Spread of the Poison Pill through the Intercorporate Network,' *Administrative Science Quarterly*, 36(4), 583-613.

Davis, G.F., Yoo, M. and W.E. Baker (2003) 'The small world of the American corporate elite, 1982-2001,' *Strategic Organization*, 1(3), 301-326.

Fichtner, J., Heemskerk, E.M., and Garcia-Bernardo, J. (2016). Hidden Power of the Big Three? Passive Index Funds, Re-Concentration of Corporate Ownership, and New Financial Risk. Available at SSRN: https://ssrn.com/abstract=2798653

Fisher, D., DeLine, R., Czerwinski, M., Drucker, S. (2012) 'Interactions with big data analytics,' *Interactions*, 19(3), 50-59.

Garcia-Bernardo, J. and Takes, F.W. (2016) `The Effects of Data Quality on the Analysis of Corporate Board Interlock Networks', *arXiv* preprint 1612.01510, 8 pages.

Glavic, Boris (2014) 'Big Data Provenance: Challenges and Implications for Benchmarking,' in T. Rabl, M. Poess, C. Baru and H-A Jacobsen (Eds) *Specifying Big Data Benchmarks*, Springer: Heidelberg, 72-80.

Haberly, D. and D. Wojcik, D. (2015). 'Earth Incorporated: Centralization and Variegation in the Global Company Network' Available at SSRN : https://ssrn.com/abstract=2699326.

Handcock, M.S., and K.J. Gile (2010) 'Modeling social networks from sampled data,' *The Annals of Applied Statistics* 4(1): 5-25.

Haunschild, P. R. (1993) 'Interorganizational Imitation: The Impact of Interlocks on Corporate Acquisition Activity,' *Administrative Science Quarterly*, 38 (4), 564-592.

Heemskerk, E.M., Daolio, G., and Tomassini, M. (2013) 'The Community Structure of the European Network of Interlocking Directorates 2005-2010,' *PloS One*, 8(7).

Heemskerk, E.M, and F.W. Takes (2016) 'The Corporate Elite Community Structure of Global Capitalism' *New Political Economy*, 21(1). 90-118.

Heemskerk, E.M., F.W. Takes, J. Garcia-Bernardo and M.J. Huijzer. 2016. 'Where is the global corporate elite? A large-scale network study of local and nonlocal interlocking directorates',*Sociologica* forthcoming. Available at arXiv:1604.04722

Henriksen, L. F., Ellersgaard, C. H., Larsen, A. G. (2016) 'Stability and change in corporate governance networks,' *paper presented at the INSNA Sunbelt conference*, April.

Jagadish, H.V. (2015) 'Big Data Science: Myths and Reality,' *Big Data Research*, 2(2), 49-52.

Kim, J. W., Kogut, B., & Yang, J. S. (2015). 'Executive compensation, Fat Cats, and best athletes' *American Sociological Review* 80(2), 299-328.

Kentor, J., & Jang, Y. S. (2004). 'Yes, there is a (growing) transnational business community a study of global interlocking directorates 1983–98' *International Sociology*, 19(3), 355-368.

Kentor, J., & Jang, Y. S. (2006). 'Different Questions, Different Answers: A Rejoiner to Carroll and Fennema' *International Sociology*, 21(4), 602–606.

Kossinets, G. (2006) 'Effects of missing data in social networks,' *Social Networks*, 28(3), 247-268.

Laney, D. (2001) '3D management: Controlling data volume, velocity, and variety,' *Application Delivery Strategies*, META Group.

Laumann, E.O., Marsden, P.V., Prensky, D. (1989) 'The boundary specification problem in network analysis,' in L. C. Freeman, D. R. White, and A. K. Romney (Eds.), *Research methods in social network analysis*. Fairfax, VA: George Mason University Press: 61-87.

Marple, T., Desmarais, B and Young, K. (2017) "Big Data, Big Corporations and Big Entity Resolution", *Unpublished Manuscript.*

Marsden, P.V. (1990) 'Network data and measurement,' *Annual Review of Sociology* 16, 435-463.

Mayer-Schönberger, V. and K. Cukier (2013) *Big Data: A revolution that will transform how we live, work, and think*, Boston : Mariner Books, Houghton Mifflin Harcourt.

Mestres, J., Gregori-Puijané, E., Valverde, S. and R. Solé (2008) 'Data Completeness – The Achilles Heel of Drug-Target Networks,' *Nature Biotechnology*, 26(9), 983-984.

Mizruchi, M.S. (1989) 'Similarity of Political Behavior Among Large American Corporations,' *American Journal of Sociology*, 95(2), 401-424.

Mizruchi, M.S. (1996) 'What Do Interlocks Do? An Analysis, Critique, and Assessment of Research on Interlocking Directorates,' *Annual Review of Sociology*, 22 (1), 271–98.

Murray, J. (2014) 'Evidence of a transnational capitalist class-for-itself: the determinants of PAC activity among foreign firms in the Global Fortune 500, 2000–2006,' *Global Networks*, 14(2), 230-250.

Neal, Z. (2014). 'The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors' *Social Networks*, *39*, 84-97.

Rao, H. and K. Sivakumar (1999) 'Institutional Sources of Boundary-spanning Structures: The Establishment of Investor Relations Departments in the Fortune 500 Industrials,' *Organization Science*, 10(1), 27–42.

Rubin, D. (1976) 'Inference and Missing Data,' *Biometrika*, 63(3), 581–92.

Rubin, D. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Salton, G. Wong, A. and Yang, C.S. (1975) "A Vector Space Model for Automatic Indexing", *Communication of the ACM* 18(11): 516-620.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., White, D. R. (2009) 'Economic networks: the new challenges' *Science*, 325 (5939), 422-5.

Squartini, T., van Lelyveld, I., & Garlaschelli, D. (2013). Early-warning signals of topological collapse in interbank networks. Scientific Reports, 3, 3357.

Starrs, S. (2013) 'American Power Hasn't Declined – It Globalized! Summoning the Data and Taking Globalization Seriously,' *International Studies Quarterly* 57(4), 817-830.

Stokman, F.N., Ziegler, R., and Scott, J. (1985) *Networks of Corporate Power*, Cambridge: Polity Press.

Takes, F.W. and Heemskerk, E.M (2016) 'Centrality in the Global Network of Corporate Control' *Social Network Analysis and Mining*, 6(1). 1-18.

Tuschke, A., Sanders, W.G., and Hernandez, E. (2014) 'Whose experience matters in the boardroom? The effects of experiential and vicarious learning on emerging market entry,' *Strategic Management Journal*, 35(3), 398-418.

Useem, M. (1984) *The inner circle: Large corporations and the rise of business political activity in the US and UK*, New York: Oxford University Press.

Van der Pijl, K., Holman, O. and O. Raviv (2011) 'The resurgence of German capital in Europe: EU integration and the restructuring of Atlantic networks of interlocking directorates after 1991,' *Review of International Political Economy,* 18(3), 384-408.

Vitali, S., Glattfelder, J. B., Battiston, S. (2011) 'The network of global corporate control,' *PloS One,* 6(10).

Wang C., Butts C.T., Hipp J.R., Jose R., Lakon C.M. (2016) 'Multiple Imputation for Missing Edge Data: A Predictive Evaluation Method with Application to Add Health,' *Social Networks* 45: 89-98.

Ward, M.D., Hoff, P.D., and Lofdahl, C.L. (2003) 'Identifying International Networks: Latent Spaces and Imputation,' in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, 345-359, Ronald Breiger, Kathleen Carley, and Philippa Pattison (eds.). Washington, D.C., The National Academies Press.

Ward, J. S., and A. Barker (2013) 'Undefined by data: a survey of big data definitions,' *arXiv,* preprint arXiv:1309.5821.

Winkler, W. E. (1990) 'String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage,' *Proceedings of the Section on Survey Research Methods (American Statistical Association)*: 354–359.

Young, K, Marple, T. and Heilman, J. (2017) "Beyond the Revolving Door: Advocacy Behavior and Social Distance to Financial Regulators", *Business and Politics*, forthcoming.

Žnidaršič, A., A. Ferligoj, and P. Doreian (2012) 'Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels,' *Social Networks* 34(4): 438-450.

Table 1: *The V's of big data mapped to problems in corporate network analysis*.

	Unit of analysis	Entity ambiguity	Completeness	Accuracy
Volume	✓	✓	✓	✓
Variety		✓	✓	
Velocity				✓
Veracity			✓	✓
Variability	✓			

Table 2. *BCND Diagnostic Toolkit*

Decision Step	**Current Diagnostic Tools**	**Future Development**
Unit of Analysis	Degree and edge distributions Core-periphery analysis Cohesiveness analysis (e.g. centralisation, cluster coefficients, coreness) Dynamic k-core decomposition	Sector, country and company type norms for degree and edge distributions, core-periphery structures and cohesiveness.
Entity Ambiguity	String matching algorithms Corporate hierarchy length Utilizing network structure such as edge multiplexity distribution, community clustering	Standardized corporate entity matching algorithms Standardized confidence intervals Standardized unique identifiers Sector, country and company type norms for corporate hierarchy.
Data Completeness	Degree and edge distributions Stratified data comparisons with known distributions. Financial variable correlations. Manual sampling and checking of edge completeness.	Sector, country and company type norms for node, edge and financial variable distributions.
Data Accuracy	Degree and edge distributions Outlier frequency	Sector, country and company type norms for node and edge distributions.

Figure 1: Decision tree for Big Corporate Network Data preparation

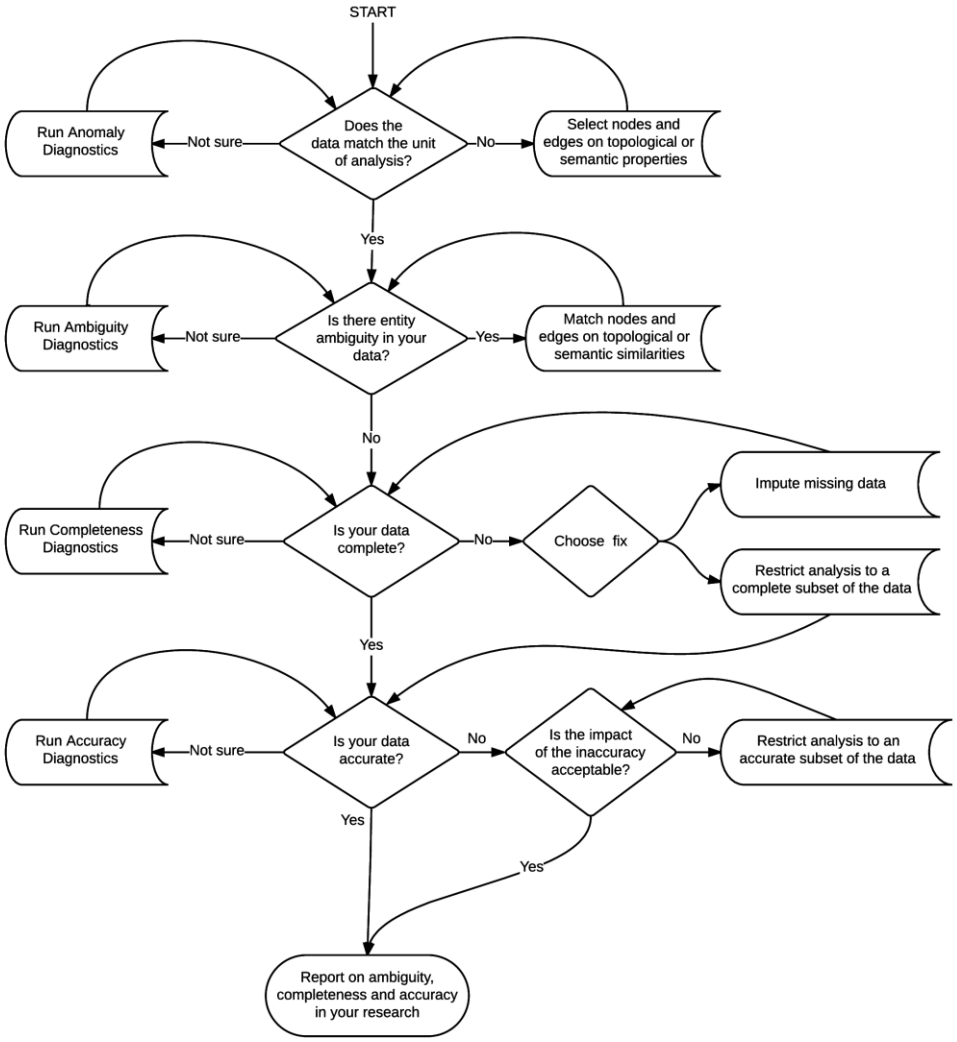


Figure 2: Number of boards before and after betweeness decomposition (a) and core size and coreness score (b) of the same data

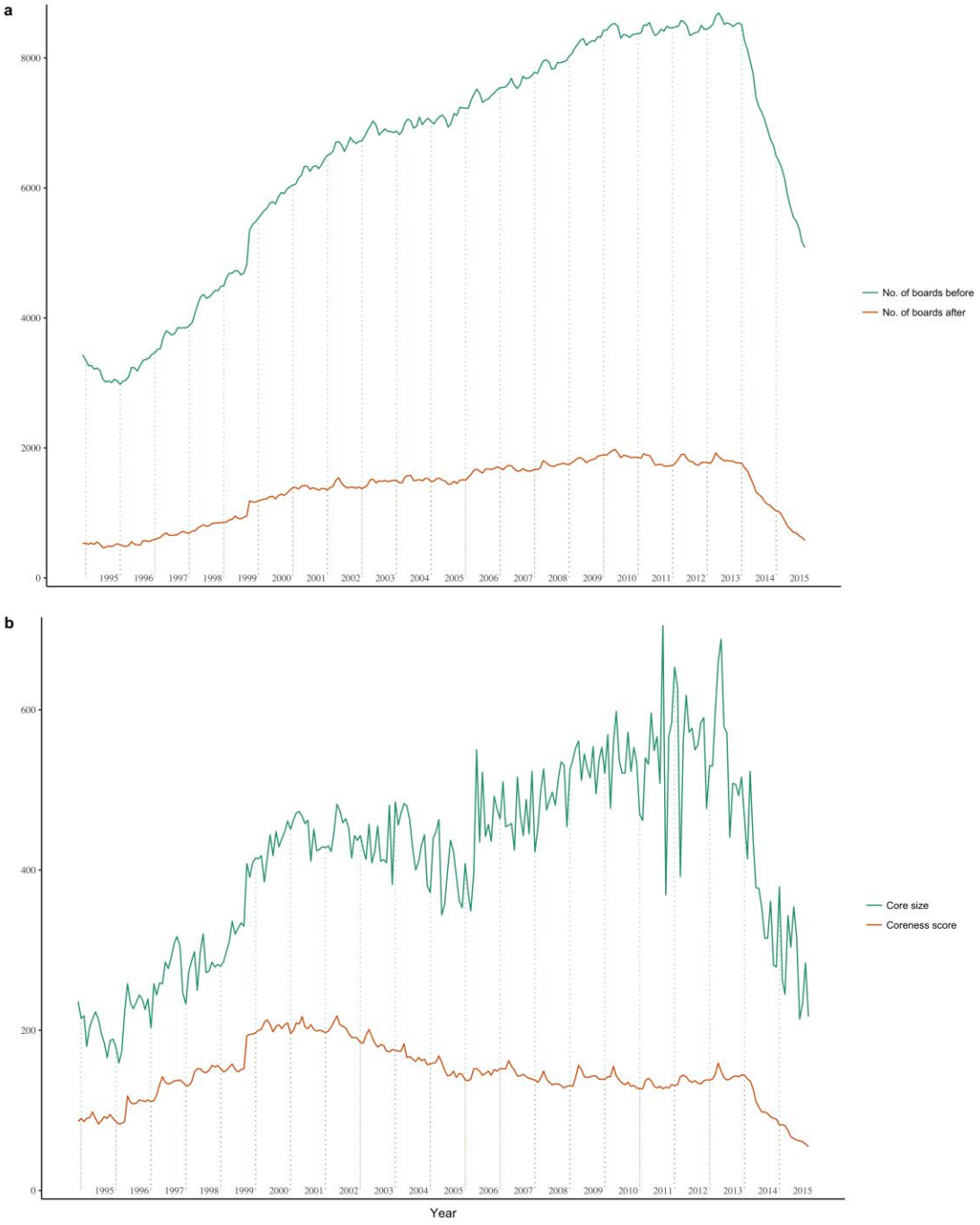


Figure 3: Goldman Sachs within ego networks of the Bank of England, US Securities and Exchange Commission (SEC) and US Federal Reserve Board

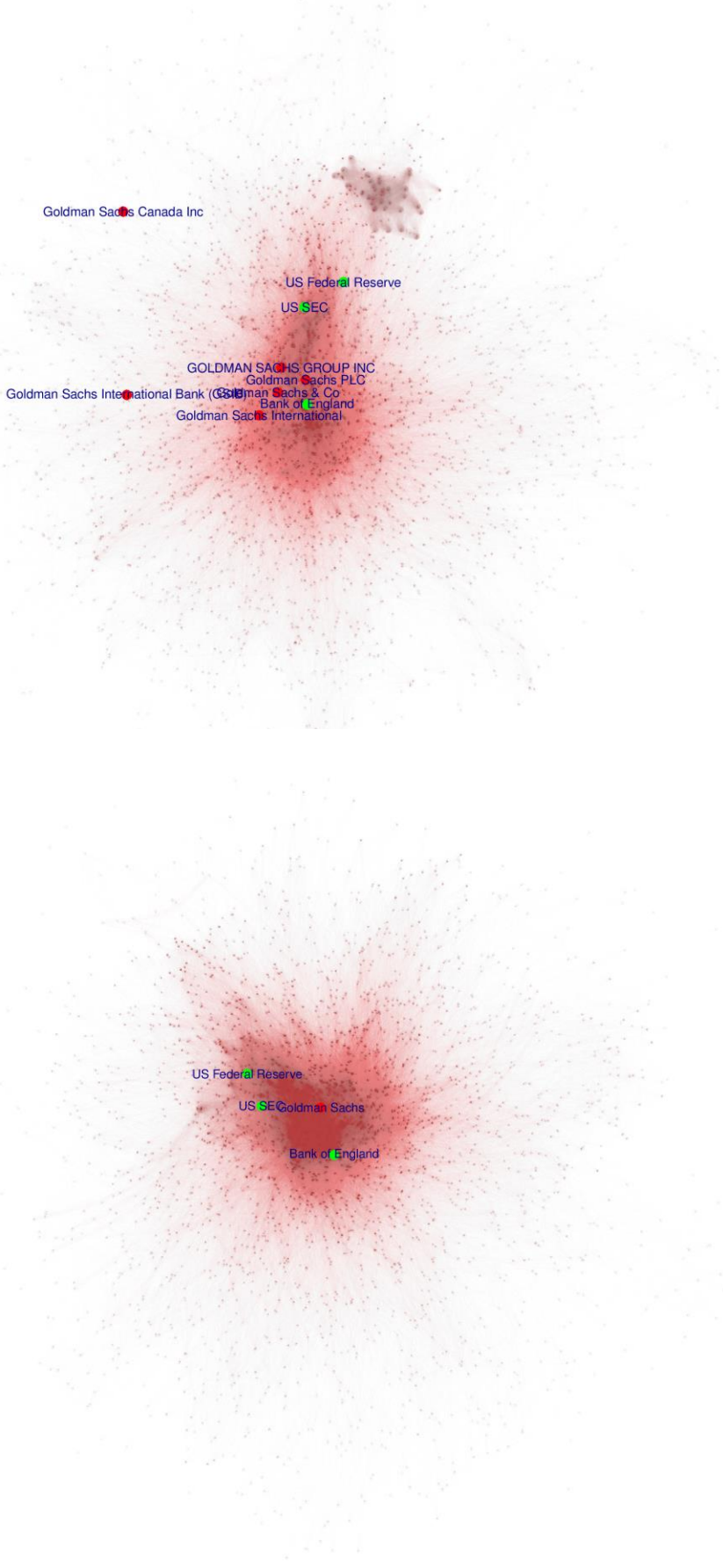


Figure 4: Two Corporate Hierarchies: Citigroup (right) and Exxon Mobil (left)

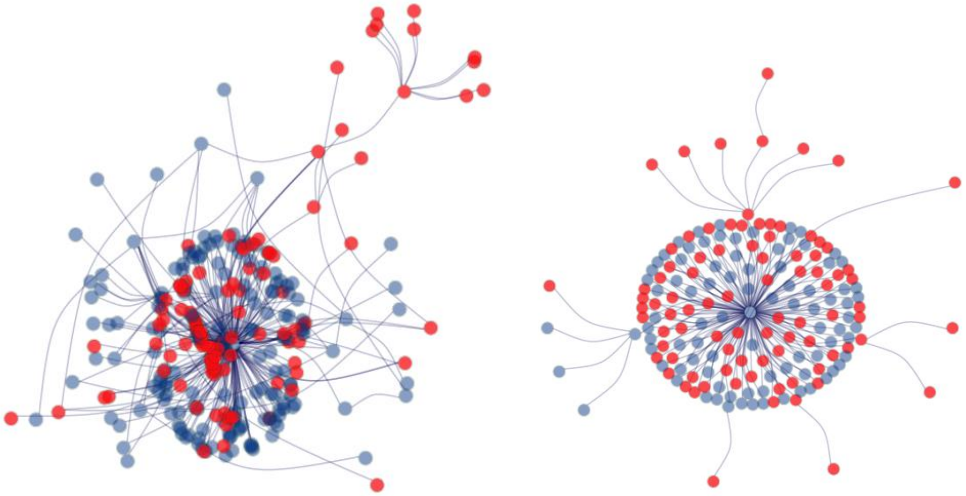


Figure 5: Number (ln) of Subsidiaries related to Global Parent (left) and Number of Total Entities in Corporate Hierarchy (right) for largest corporations in the world

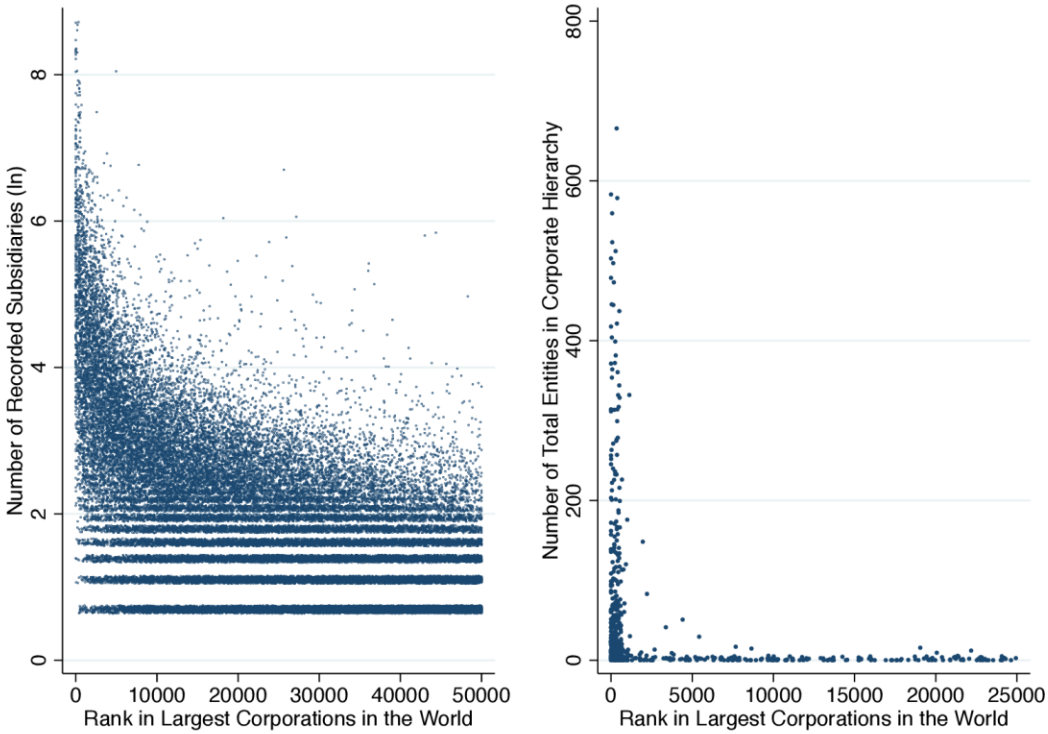
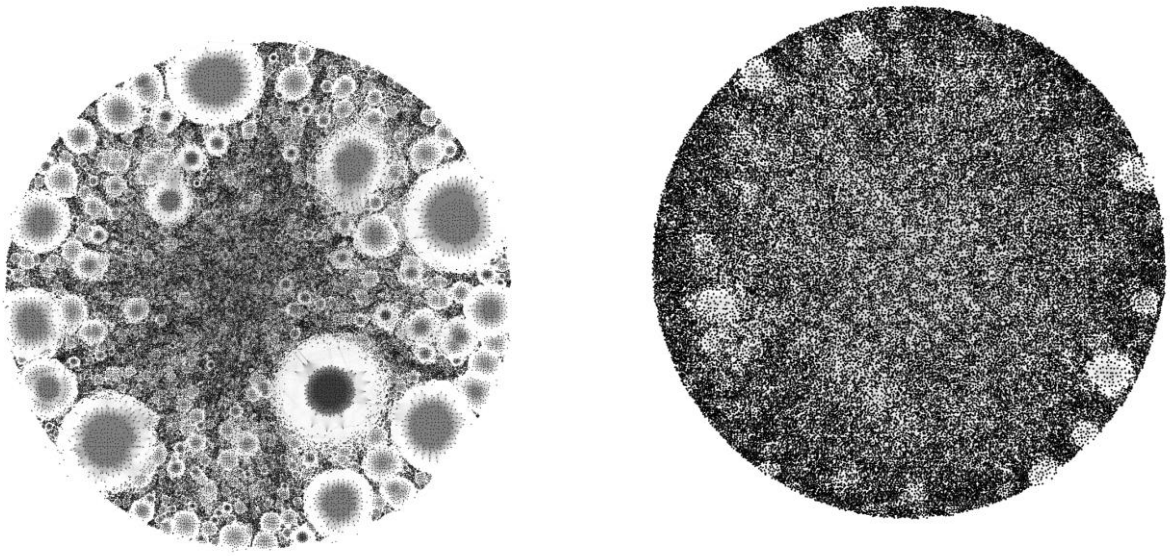
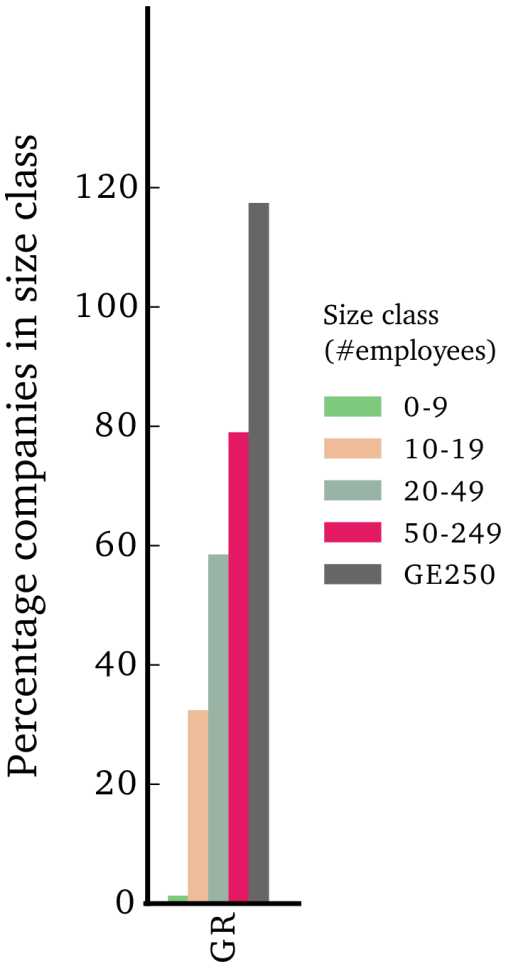


Figure 6 Entity ambiguity can create artifacts in the data



Network visualizations of the Swedish interlock network using the ForceAtlas2 algorithm for the raw network (left) and the corrected network (right) using the method described in Garcia-Bernardo and Takes (2016)

Figure 7: Datasets are not equally complete for small and large companies



The percentage of companies present in Orbis in comparison with the Eurostat database as a function of the number of employees. GE250 denotes more than 250 employees.

Figure 8: *The distribution of the board size in Panama*

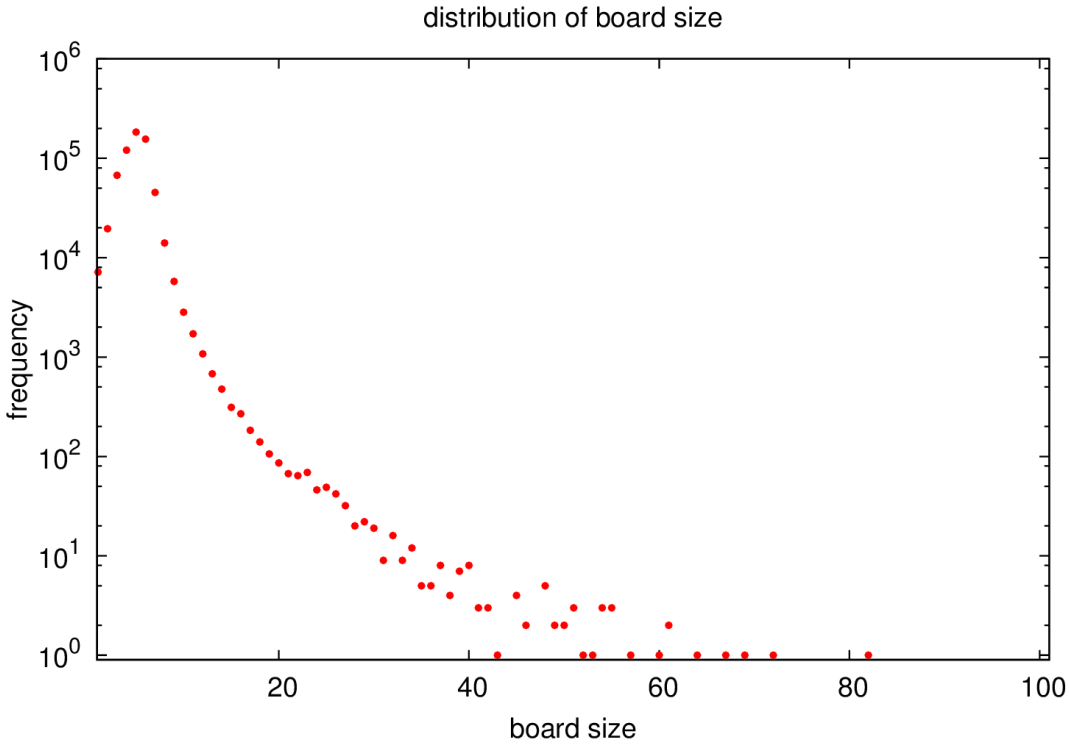


Figure 9: *The distribution of the number of director positions in five different countries*